PROJECT REPORT

EFFICIENT WATER QUALITY ANALYSIS AND PREDICTION USING MACHINE LEARNING

DONE BY

TEAM ID: PNT2022TMID46440

KALANITHI M(820319205014)
BHUVANESHWAR S(820319205008)
KARTHICKRAJA R(820319205016)
GIRIDHARAN N(820319205301)

GUIDED BY

B.N.KARTHIK (MENTOR)

Ms. LALITHA GAYATHRI (INDUSTRY MENTOR)

INDEX

7. **CODING & SOLUTIONING (Explain the features added in the project along with code)**
    a. Feature 1
    b. Feature 2
    c. Database Schema (if Applicable)
8. **TESTING**
    a. Test Cases
    b. User Acceptance Testing
9. **RESULTS**
    a. Performance Metrics
10. **ADVANTAGES& DISADVANTAGES**
11. **CONCLUSION**
12. **FUTURE SCOPE**
13. **APPENDIX**
    Source Code
    GitHub & Project Demo Link

EFFICIENT WATER QUALITY ANALYSIS ANDPREDICTION USING MACHINE LEARNING

1. INTRODUCTION

a. PROJECT OVERVIEW

Water is considered as a vital resource that affects various aspects of human health and lives. The quality of water is a major concern for people living in urban areas. The quality of water serves as a powerful environmental determinant and a foundation for the prevention and control of waterborne diseases. However predicting the urban water quality is a challenging task since the water quality varies in urban spaces non-linearly and depends on multiple factors, such as meteorology, water usage patterns, and land uses, so this project aims at building a Machine Learning (ML) model to Predict Water Quality by considering all water quality standard indicators. Water is the most important of sources, vital for sustaining all kinds of life; however, it is in constant threat of pollution by life itself. Water is one of the most communicable mediums with a far reach. Rapid industrialization has consequently led to deterioration of water quality at an alarming rate. Poor water quality results have been known to be one of them factors of escalation of harrowing diseases. As reported, in developing countries, 80% of the diseases are water borne diseases, which have led to 5 million deaths and 2.5 billion illnesses. The most common of these diseases in Pakistan are diarrhoea, typhoid gastroenteritis, cryptosporidium infections, some forms of hepatitis and giardiasis intestinal worms. In Pakistan, water borne diseases, cause a GDP loss of 0.6–1.44% every year. This makes it a pressing problem, particularly in a developing country like Pakistan. Water quality is currently estimated through expensive and time-consuming lab and statistical analyses, which require sample collection, transport to labs, and a considerable amount of time and calculation, which is quite ineffective given water is quite a communicable medium and time is of the essence if water is polluted with disease-inducing waste.

The horrific consequences of water pollution necessitate a quicker and cheaper alternative. In this regard, the main motivation in this study is to propose and evaluate an alternative method based on supervised machine learning for the efficient prediction of water quality in real-time. A representative set of supervised machine learning algorithms were employed on the said dataset for predicting the water quality index (WQI) and water quality class (WQC). The main contributions of this study are summarized as follows. A first analysis was conducted on the available data to clean, normalize and perform feature selection on the water quality measures, and therefore, to obtain the minimum relevant subset that allows high precision with low cost. In this way, expensive and cumbersome lab analysis with specific sensors can be avoided in further similar analyses. A series of representative supervised prediction (classification and regression) algorithms were tested on the dataset worked here. The complete methodology is proposed in the context of water quality numerical analysis.

b. PURPOSE

Water makes up about 70% of the earth's surface and is one of the most important sources vital to sustaining life. Rapid urbanization and industrialization have led to a deterioration of water quality at an alarming rate, resulting in harrowing diseases. Water quality has been conventionally estimated through expensive and time-consuming lab and statistical analyses, which render the contemporary notion of real-time monitoring moot. The alarming consequences of poor water quality necessitate an alternative method, which is quicker and inexpensive. With this motivation, this research explores a series of supervised machine learning algorithms to estimate the water quality index (WQI), which is a singular index to describe the general quality of water, and the water quality class (WQC), which is a distinctiveclass defined on the basis of the WQI.

The proposed methodology employs four input parameters, namely, temperature, turbidity, Ph and total dissolved solids. Of all the employed algorithms, gradient boosting, with a learning rate of 0.1 and polynomial regression, with a degree of 2, predict the WQI most efficiently, having a mean absolute error (MAE) of 1.9642 and 2.7273, respectively. Whereas multi-layer perceptron (MLP), with a configuration of (3, 7), classifies the WQC most efficiently, with an accuracy of 0.8507. The proposed methodology achieves reasonable accuracy using a minimal number of parameters to validate the possibility of its use in real time water quality detection systems.

2. LITERATURE REVIEW

a. EXISTING PROBLEM

The basic idea of this research is to devise a comprehensive methodology that analyzes and predicts the water quality of particular regions with the help of certain water quality parameters. These parameters include physical, biological, or chemical factors which influence water quality. There are certain quality standards set up by international organizations like the World Health Organization (WHO) and the Environmental Protection Agency (EPA), which serve as a benchmark for determining the quality of water. In its document “Efficient Water Quality Analysis and Prediction using Machine Learning”, EPA mentions a total of 101 parameters that affect water quality in one way or another. However, some parameters have a greater and more visible effect on water quality than others.

TITLE: IMPROVING THE ROBUSTNESS OF BEACH WATER QUALITY MODELING USINGAN ENSEMBLE MACHINELEARNING

AUTHOR: Wang et al (2021)

This study demonstrates the utility of using a model stacking approach for predictive modeling of beach water quality. Since model stacking averages out noise from its base models, it is theoretically more promising than individual models in generating predictions with greater accuracy and robustness. The results from this study suggest that the model stacking algorithm has promise for improving the reliability of predictive modeling for beach microbial water quality of other sites with similar hydrogeological and environmental conditions such as other beaches along the Great Lakes. A comprehensive test needs to be done to understand the strength a d weaknesses of individual base models and the stacking approach. This study indicated that the model stacking approach may improve the robustness of beach water quality modeling.

TITLE: ACCURATE PREDICTION SCHEME OF WATERQUALITY IN SMARTMARICULTURE WITH A DEEP BI-S-SRU LEARNING NETWORK

AUTHOR: J. Liu, C. Yu, Z. Hu et al (2020)

This paper proposed the process and model for the accurate prediction of key water quality parameters (pH, water temperature, and dissolved oxygen). Firstly, the collected water quality data is repaired and corrected by the improved preprocessing method, and then the data is filtered and noised by the wavelet transform method. After preprocessing, the data received by remote transmission can be recovered well. Next, we construct the Bi-S-SRU (Bi-directional Stacked SRU) deep learning prediction model by importing a pretreated dataset weighted with the discovered correlation coefficients. The experimental results demonstrate that our proposed prediction model can achieve higher prediction accuracy and stability compared with RNN-based and SRU-based prediction models. The experimental results also show that the Bi-S-SRU-based prediction method is only slightly higher in time complexity than the traditional RNN-basedor LSTM-based prediction method.

TITLE: ASSESSMENT OF SURFACEWATER QUALITY BY USING SATELLITEIMAGES FUSION BASED ON PCA METHOD IN THE LAKE GALA, TURKEY

AUTHOR: E. Batur and D. Makita (2019)

In this paper, the PCA model is presented to integrate surface water reflectance values from satellite images to monitor Gala Lake's surface water quality. The values of Chl-a, DO, TSS, SDD, TDS, and pH values calculated by the PCA method were found to be highly correlated with the measured water quality parameters. The results obtained were found to be directly proportional to the number of sensors. L8 OLI and S2A have higher spectral resolution than GK2 images. However, the high temporal resolution of GK2allows the desired region to be displayed at more frequent intervals, allowing for better monitoring of the instantaneous changes in surface water quality. Therefore, longer measurements should be made and analyzed for a model covering all periods.

TITLE: SURFACE WATER POLLUTIONDETECTION USING THE INTERNET OF THINGS

AUTHOR: Shafietal (2018)

In this paper, the proposed an IoT-based solution to monitor water quality in real-time. The proposed system provides remote monitoring of water quality assessment along with water flow control via a mobile app. Four machine learning algorithms including Support Vector Machine (SVM), k Nearest Neighbor (kNN), single layer neural network, and deep neural network have been applied for the classification of water quality and experimental results revealed that deep neural network outperforms all other algorithms with an accuracy of 93. This system has the potential to effectively utilize to overcome the challenges of water quality in the agriculture sector and various industries.

TITLE: IMPROVING WATER QUALITY INDEX PREDICTION IN PERAK RIVER BASIN MALAYSIA THROUGH A COMBINATION OF MULTIPLE NEURAL NETWORKS

AUTHOR: Ahmad et al (2017)

In this paper, they proposed a reliable real-time prediction model for WQI developed through a selective combination of multiple neural networks by excluding COD and BOD from model inputs as they cannot be measured in real-time. Single and multiple FANN are used in this paper to model the WQI in the Perak River basin. The selective combination schemes provide models with better generalization capability compared to combining all neural networks. The bootstrap aggregated models with selective combination provide a real-time WQI prediction tool without delay as holy real-time measurement are used as model inputs.

TITLE: ARTIFICIAL INTELLIGENCE FOR THE PREDICTION OF WATER QUALITY INDEX IN GROUND WATER SYSTEMS

AUTHOR: Mohamad Sakizadeh (2016)

One of the problems of ANN's modeling in environmental studies which suffers from the problem of the small data records is the danger of over-fitting the model to the training data resulting in poor generalization of the model for the data out-of-the training data range. This study's results proved that this problem can be obviated by using some algorithms like Bayesian regularization and Ensemble methods. The prediction of water quality index (WQI) was successfully implemented by Bayesian regularization and Ensemble averaging methods, though the performance of Bayesian regularization was roughly better, with minimum test error indicating the good generalization ability of these methods in this field. The poor generalization ability is a problem that has been overlooked by most of the research all around the world although it is an important issue that should be taken into account.

TITLE: THE USE OF COMBINEDNEURAL NETWORKS AND GENETIC ALGORITHMS FORTHE PREDICTION OF RIVER WATER QUALITY

AUTHOR: Ding et al (2014)

In this paper, they propose a water quality prediction model that combines PCA, BPNN, and GA. Using the BPNN model to study water classification and prediction can overcome disadvantages including the large workload of traditional evaluation methods and strong subjectivity. This model possesses objectivity, universality, and practicality. PCA converts the multi-indices into a few aggregative indices with little original data information loss and reduces the input data to speed the training process. Using GA to optimize network parameters can effectively prevent the search process rom converging to local optimum solutions, optimize global optimal network parameters, and significantly improve the accuracy of water quality prediction. This model can obtain high training speed and good prediction rate and can be extended to other classification problems.

Our Ideology

The estimated water quality in our work is based on nine parameters: ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic carbon, Trihalome thanes, Turbidity, and pH, which are tested according to World Health Organization (WHO) standards. The proposed

methodology improves on these notion sand the methodology being followed is depicted in Figure..

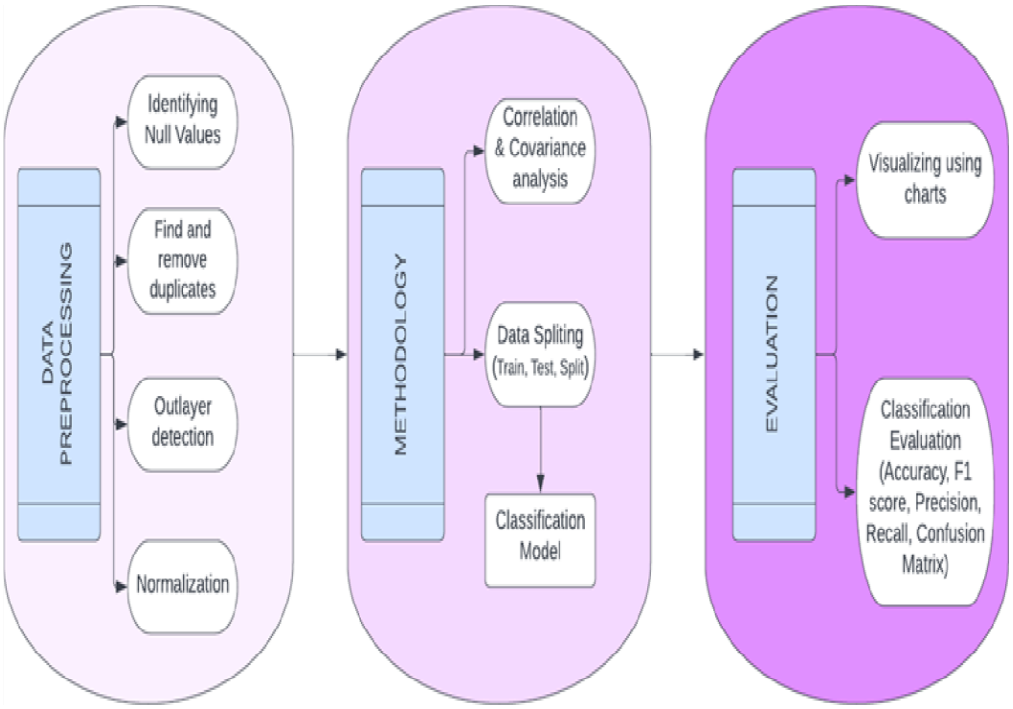


Figure 1

COMPARATIVE ANALYSIS OF LITERATURE SURVEY:

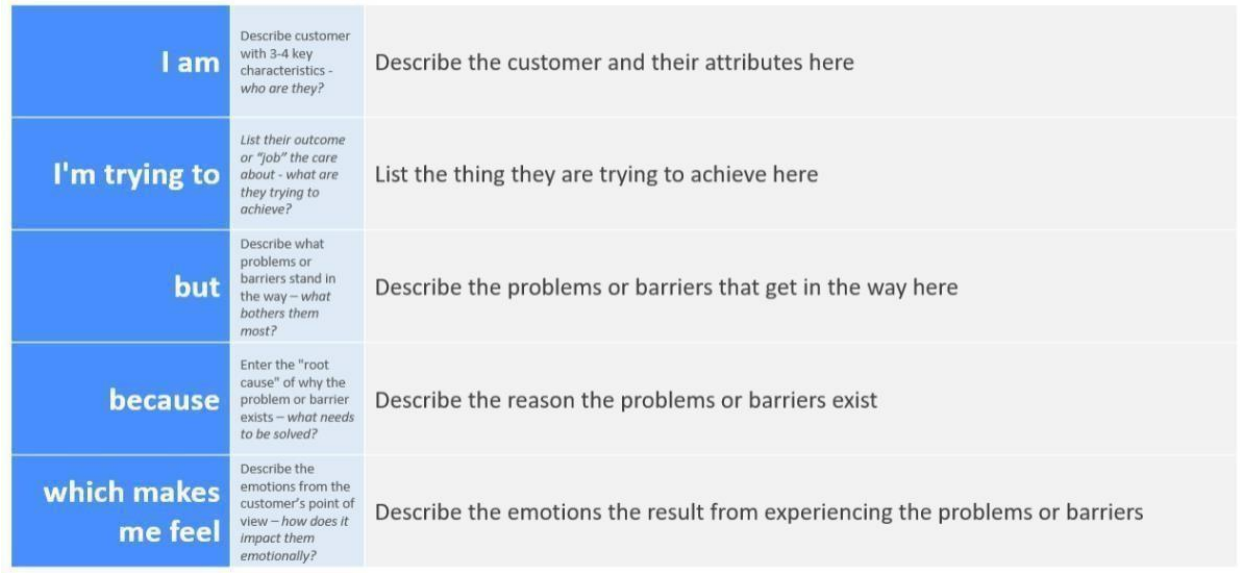
AUTHOR	TITLE	YEAR	SOURCE	FINDINGS	PROS	CONS
[1] Yafra Khan, Chai Soo See, Min,Zeilh ofer, FarrellPoe, Cernadas, Barro, Amorim	Predicting and analyzing water quality using Machine Learning	2016	International Conference on Recent Trends on Electronics, Information, Communicati on & Technology (RTEICT).	Artificial Neural Networks, Time Series Analysis, Mean Square Error Methods, Predictive Models	Most significantly utilized to predict the water in order to avoid filthiest.	The Quality of water might not be accurate sometime.
[2] Rongli Gai, Jiahui Yang, Nikhil M Ragi, Ravishankar Holla, Manju, Pedregosa, Varoquaux, Gramfort, Michel	Predicting water quality parameter s using machine learning	2019	International Conference on Digital Information Management (ICDIM).	Chemical methods, Predicting chloride, neural nets, Biological neural networks, Classification algorithms.	With the assist of, Water Quality Prediction technique, could be ignored impecunious people death for the reason of, utilizing corporation water.	If foretelling the water using chemical methods in pure natural water like, Filtered Rain Water then, it would cause some menacing
[3] Kathleen Joslyn, John Lipor, Basak Pal, Patranabis, Chen, Guestrin	A Supervised Learning Approach toWater Quality Parameter Prediction and Fault Detection	2018	IEEE International Conference on Big Data (Big Data).	Sensors, Qfactors, Support Vector Machines, Regression Analysis, Gradient boosting, Dissolved Oxygen, Turbidity	A Supervised Learning. It has many algorithms to detect the best	It has many algorithms to detect the best quality of the water. Not only depends on Supervised Learning approach.
[4] Jitha Nair, Vijaya, Tirabassi, Liao, Ahmed, Mumtaz, Anwar, Irfan, GarciaNieto, Shah.	Predictive Models for River Water Quality using Machine Learning and Big Data Techniques.	2021	International Conference on Artificial Intelligence and Smart Systems (ICAIS).	Big Data, Remote Sensing Data analysis, Deep learning, Environmental Science Computing, Predictive Models.	It is more crucial, that foretelling the river water using machine learning and bi data techniques to defend the diseases.	Prediction must be meticulous in order to ignore bad reports about the water.

[5] MalarKodi, Tarakeswari, Jobin Tomy, Ashton, Joubert, Benyamina, Gamatie, Mahadeo, Mehmood.	A Deep Learning Strategy For Water Quality Monitoring		A Deep Learning Strategy For Water Quality Monitoring	Internet Of Things, Existing Water Quality Forecasting Techniques, LSTM, Total Dissolved Solids.	With the assist of, monitoring the water quality, it would be so effortless to ignore diseases.	Monitoring simultaneously might be arduous.

c. PROBLEM STATEMENT DEFINITION

Customer Problem Statement:

Create a problem statement to understand your customer's point of view. The Customer Problem Statement template helps you focus on what matters to create experiences people will love. A well-articulated customer problem statement allows you and your team to find the ideal solution for the challenges your customers face. Throughout the process, you'll also be able to empathize with your customers, which helps you better understand how they perceive your product or service.



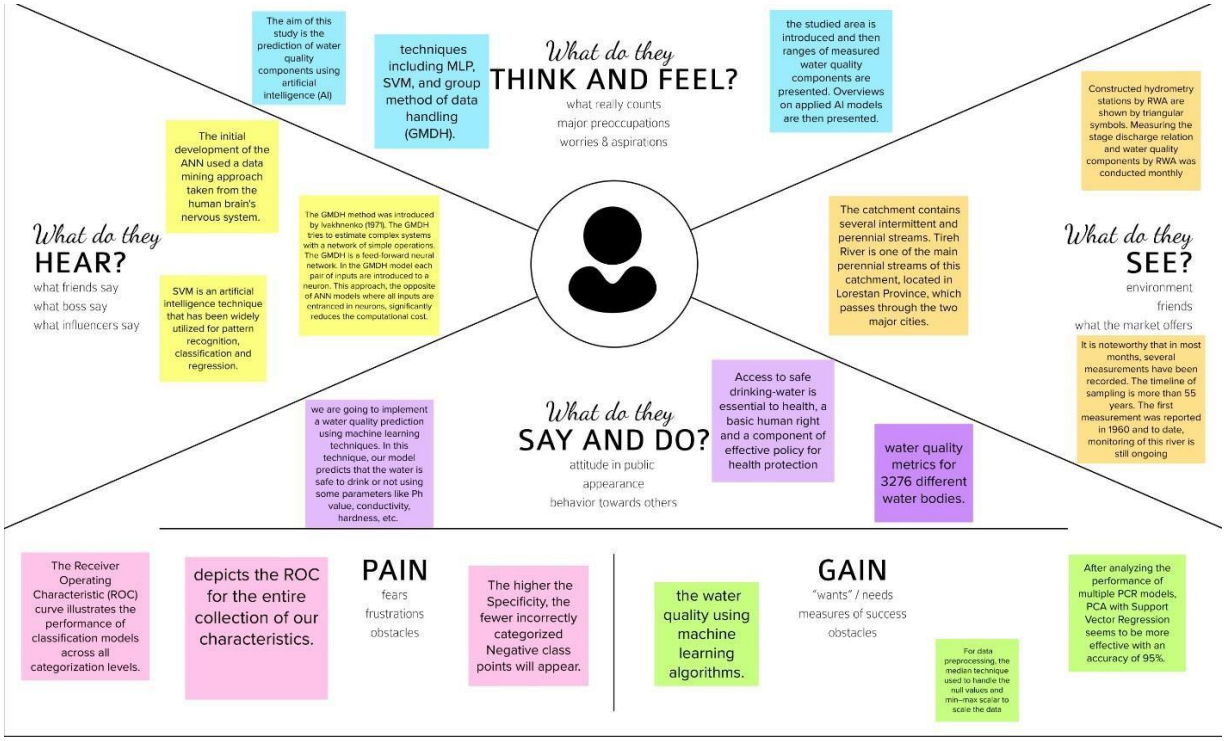
Problem Statement (PS)	**I am (Customer)**	**I'm trying to**	**But**	**Because**	**Which makes me feel**
PS – 1	Sanitization companies	Need a good quality measure to use upon their products	Unaware of how to measure the water quality	I can not find an apt metric to rely on	Twitchy
PS – 2	Farmers	Water my field with a good quality possible	Unaware of where to find it	Mostly available water is contaminated	Edgy

PS – 3	Diseased people	Find good quality water	Unaware of where to find it	Mostly available water is contaminated	Restless
PS – 4	Lab technicians	Need a qualitative model to satisfy the sanitization company needs	Unaware of water quality index model	Many models don't provide a best metric to measure the quality	Flurried
PS – 5	People who wish to lead a healthy life	Find good quality water	Unaware of where to find it	Mostly available water is contaminated	Concerned

3. IDEATION & PROPOSED SOLUTION

a. EMPATHY MAP CANVAS

Efficient Water Quality Analysis and Prediction Using Machine Learning

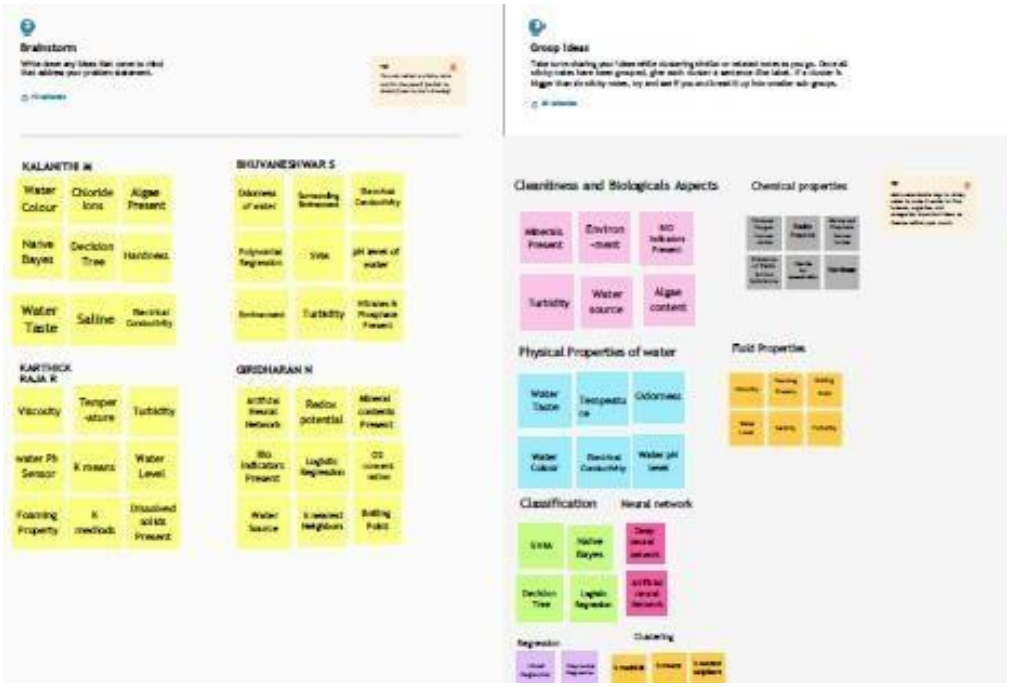


b. IDEATION & BRAINSTORMING

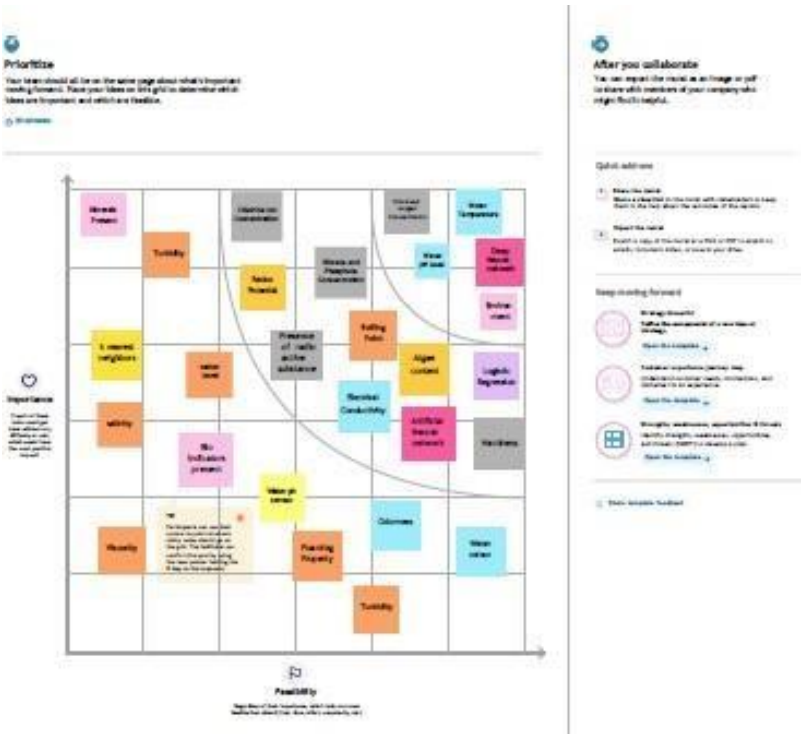
Brainstorm & Idea Prioritization Template:

Step-1: Team Gathering, Collaboration and Select the Problem Statement

Step-2: Brainstorm, Idea Listing and Grouping



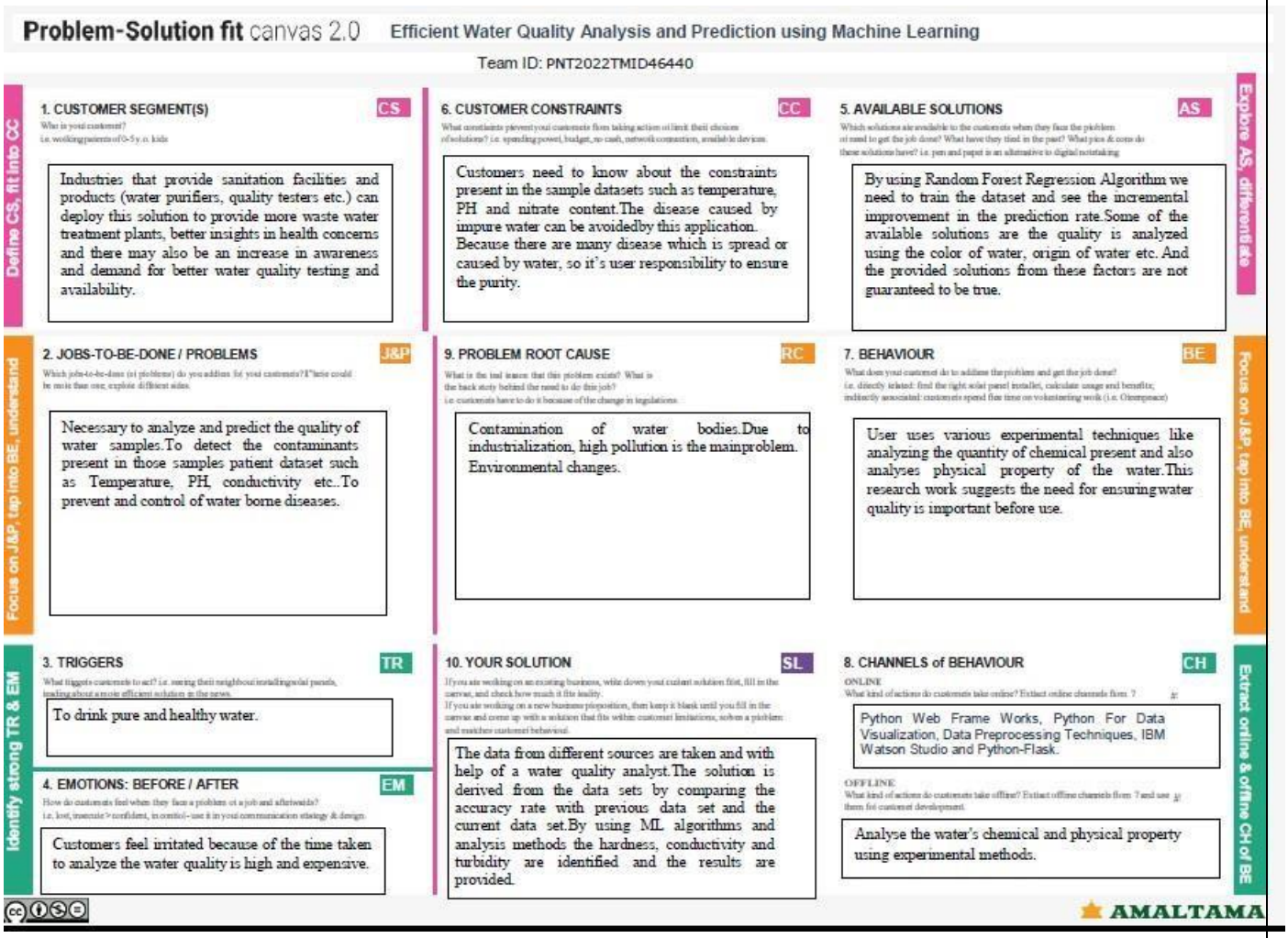
Step-3: Idea Prioritization



c. PROPOSED SOLUTION

S.NO	PARAMETER	DESCRIPTION
1.	**Problem Statement (Problem to be solved)**	Efficient Water Quality Analysis and Prediction using Machine Learning.
2.	**Idea / Solution description**	For the WQI prediction, artificial neural network models, namely nonlinear autoregressive neural network (NARNET) and long short-term memory (LSTM) deep learning algorithm, have been developed. In addition, three machine learning algorithms, namely, support vector machine (SVM), K- nearest neighbor (K-NN), and Naive Bayes, have been used for the WQC forecasting. The used dataset has 7 significant parameters, and the developed models were evaluated based on some statistical parameters
3.	**Novelty / Uniqueness**	In previous they find water quality with help of WQI and WQC. Now the solution is find with help of advanced artificial intelligence and it include seven parameters
4.	**Social Impact / Customer Satisfaction**	During the last years, water quality has been threatened by various pollutants. Therefore, modelling and predicting water quality have become very important in controlling water pollution. In this work, advanced artificial intelligence (AI) algorithms are developed to predict water quality index (WQI) and water quality classification (WQC).This is the impact of this statement
5.	**Business Model (Revenue Model)**	The revenue stream include the Promoted trends and method. Technology and production is improved in business side. It increased the profit and also the logistic way.
6.	**Scalability of the Solution**	Scalability of this solution can handle any amount of data and perform many computations in a cost effective and time saving to instantly serve millions of users residing at global location.

d. PROBLEM SOLUTION FIT



4. REQUIREMENT ANALYSIS

a . Functional Requirements:

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	Check water quality analysis	Water's quality is more important which should be considered as many water-borne diseases are more widely known. So, it is necessary to analyze and predict the quality of water samples so as to determine and detect the contaminants present in those samples Patient dataset such as Temperature, PH, Conductivity, B.O.D, Nitrate nan, Fecal Coliform, Total Coliform, Year etc
FR-2	Predict Water Quality by considering all water quality standard indicators	Using Machine learning model
FR-3	Accessing datasets	Datasets are collected by data preprocessing method then followed by data visualization
FR-4	Classification of dataset	Dataset includes of data exploration. In which prediction of water quality index calculation is performed using KNN ,SVM, ANN, Naivis bayes and linear regression algorithms
FR-5	Splitting and train the data	In this phase, we split the dataset into training and test dataset , and then trained the models using training dataset

b. Non-functional Requirements:

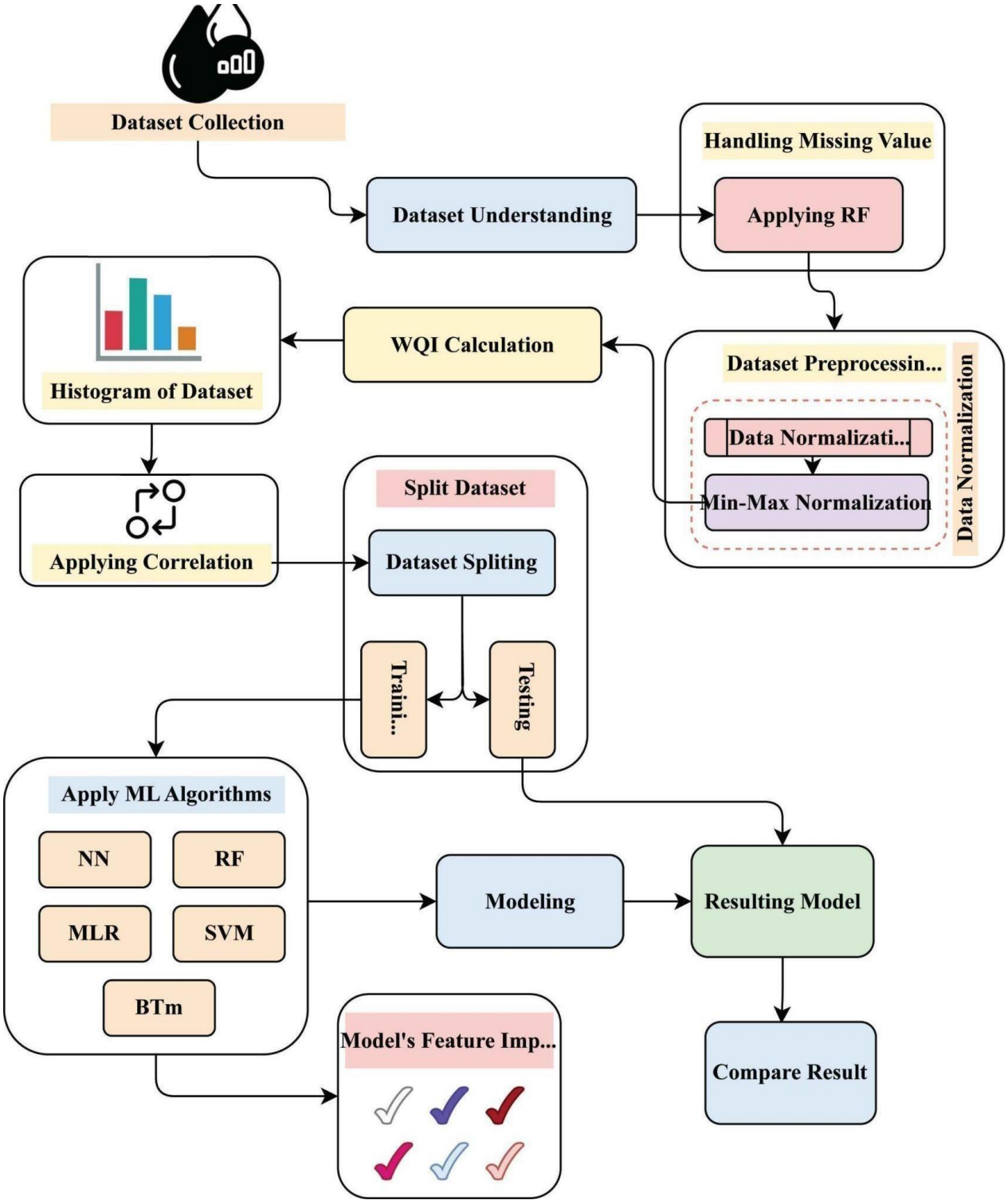
Following are the non-functional requirements of the proposed solution.

FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	Predicting the urban water quality is a challenging task since the water quality varies in urban spaces non- linearly and depends on multiple factors, such as meteorology, water usage patterns, and land uses.
NFR-2	**Security**	The quality of water is a major concern for people living in urban areas.The quality of water serves as a powerful environmental determinant and a foundation for the prevention and control of waterborne diseases
NFR-3	**Reliability**	This project will help everyone in protecting their health. Accurate water quality prediction is the basis of water environment management and is of great significance for water environment protection.
NFR-4	**Performance**	This system uses different sensors for monitoring the water quality by determine pH, Turbidity, conductivity and temperature. Data is gathered from different sources it is collected in a raw format and this data isn't feasible for the analysis

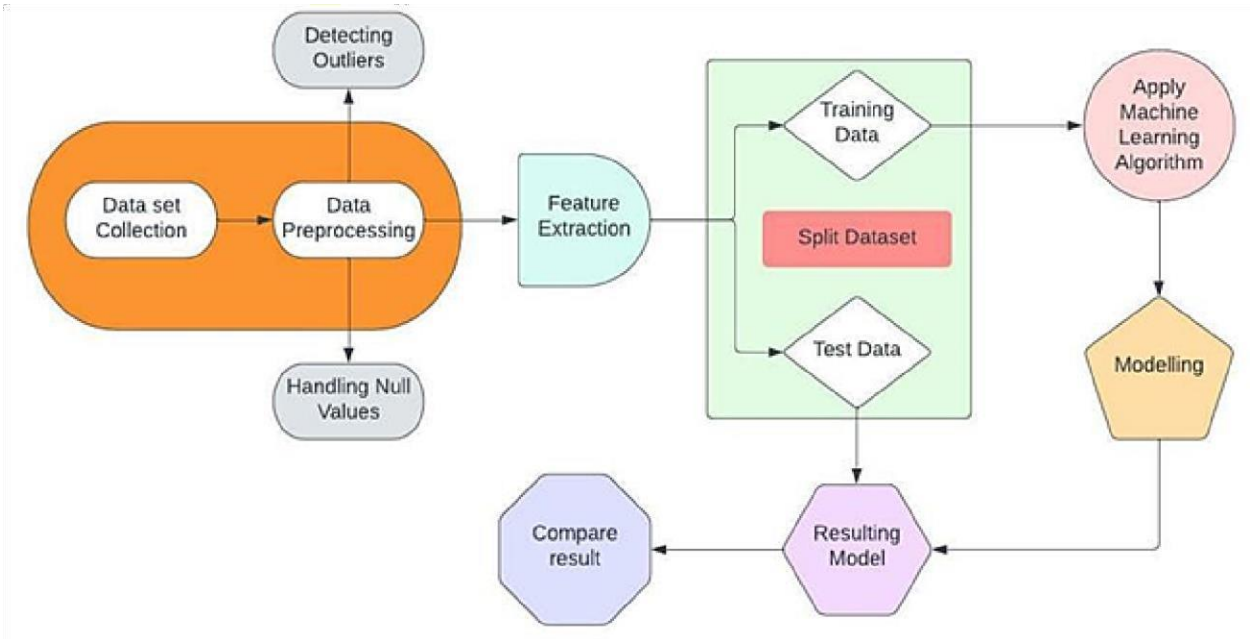
NFR-5	**Availability**	Industries that provide sanitation facilities and products (like water purifiers, quality testers etc.) can deploy this solution to provide more waste water treatment plants, better insights in health concerns and there may also be an increase in awareness and demand for better water quality testing and availability.
NFR-6	**Scalability**	This project used to measure and determine the quality of water. This provides pollution free and purified water.

5. PROJECT DESIGN

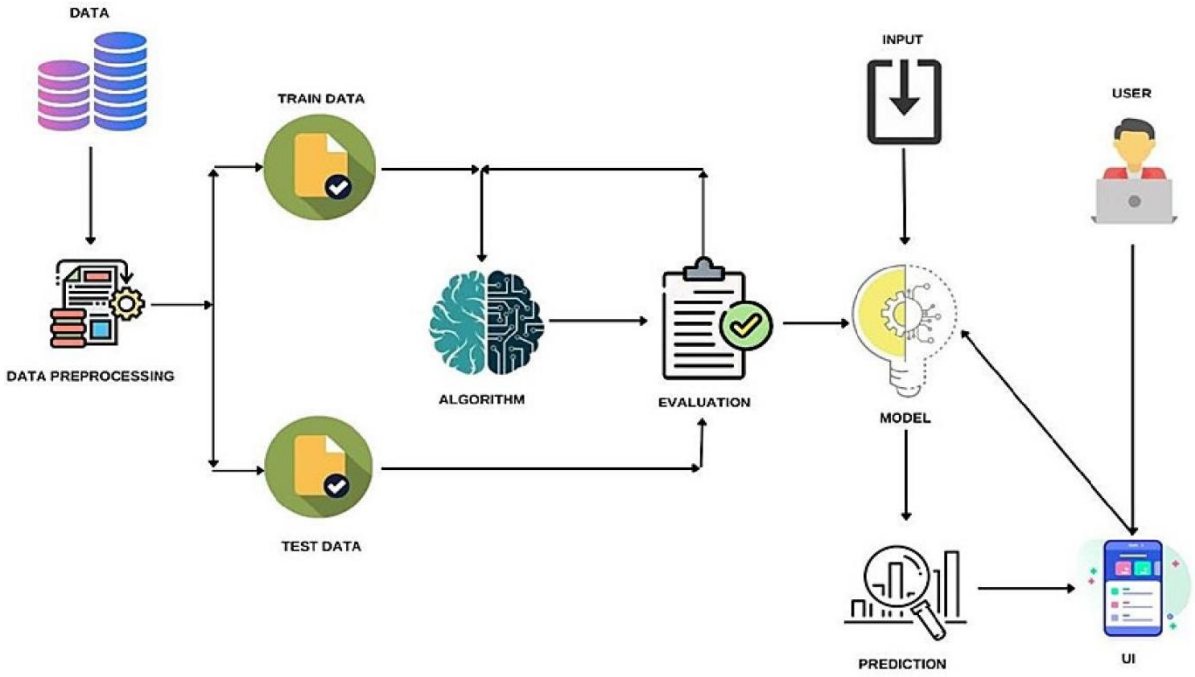
a. DATA FLOW DIAGRAMS


Dataset Collection
Dataset Understanding
Handling Missing Value
Applying RF
Histogram of Dataset
WQI Calculation
Dataset Preprocessin...
Data Normalizati...
Min-Max Normalization
Data Normalization
Applying Correlation
Split Dataset
Dataset Spliting
Traini...
Testing
Apply ML Algorithms
NN
RF
MLR
SVM
BTm
Modeling
Resulting Model
Model's Feature Imp...
Compare Result

b. SOLUTION AND TECHNICAL ARCHITECTURE



Solution Architecture



Technical Architecture

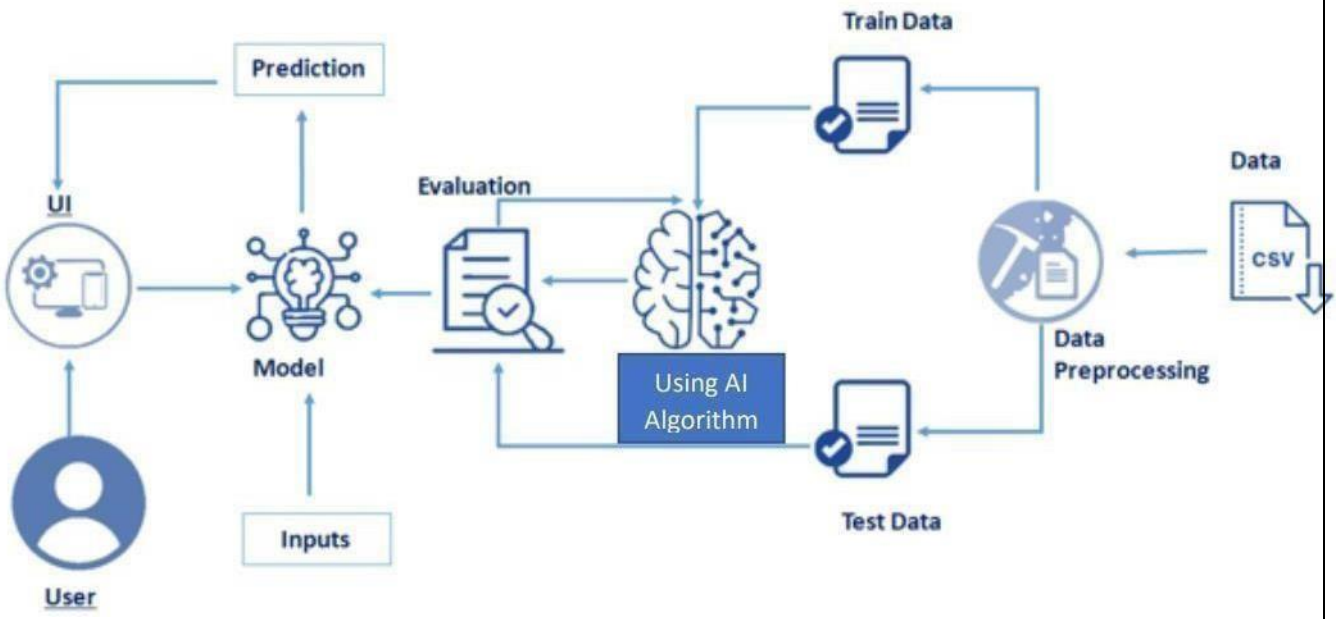


Table-1: Components & Technologies:

S. No	Component	Description	Technology
1.	User Interface	User interacts by using web user interface.	HTML, CSS and Python Flask
2.	Application Logic-1 (Login)	User can able to login if that person is already registered to the site.	HTML, CSS and Python Flask
3.	Application Logic-2 (Register)	User needs to be registered if that person is new to the site.	HTML, CSS and Python Flask.

4.	Application Logic-3(Reporting Form)	User needs to click on the reporting form in order to get the prediction result	Front end- HTML,CSS and Python Flask. Back end – Query Languages, Python.
5.	Database	Data Type-String, Numeral values.	Query Languages such as MySQL, NoSQL etc.
6.	Cloud Database	Database Service on Cloud.	IBM DB2, IBM Cloud ant etc.
7.	File Storage	File storage requirements.	Local File-system.
8.	External API-1	Anyone can access the details with some restrictions to the personal details of other users.	Web API.
9.	External API-2	Accessibility.	Aadhar API.
10.	Machine Learning Model	Predict the result based on the training and testing dataset.	Data Recognition Model, etc.
11.	Infrastructure (Server / Cloud)	Application Deployment on Local System.	Local System.

Table-2: Application Characteristics:

S. No	Characteristics	Description	Technology
1.	Open-Source Frameworks	Frameworks are used for predictive data analysis, providing clear and actionable error messages.	Tensor flow, Sci-kitlearn, Keras.
2.	Security Implementations	OTP will be sent to the registered email id. Unauthorized users could not access the user's details.	Email Verification.
3.	Scalable Architecture	Scalability is improved for implementing the three-tier architecture.	Three tier architecture.
4.	Availability	For enhancing the high availability, load balancer is needed.	Load Balancer.
5.	Performance	The model could be able to process large number of datasets.	Load Balancer.

c. USER STORIES

Here we have listed all the user stories for the product.

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Mobile user)	Usability & Compatibility	USN-1	As a mobile user, I want to use the application using my mobile phone.	I can use my phone to access the website.	High	Sprint-1
	Login & Registration	USN-2	As a user, I can register for the application through Facebook.	I can register & access the dashboard with Facebook Login	Medium	Sprint-2
	Authentication	USN-3	As a user, I will receive a confirmation email once I have registered for the application	I can receive a confirmation email & click confirm	Low	Sprint-2
Customer (Web user)	Login & Registration	USN-4	As a user, I can register for the application through Gmail	I can register & access the dashboard with a Gmail login	Medium	Sprint-1
		USN-5	As a user, I can log into the application by entering my email & password	I can get to the dashboard after signing in	Medium	Sprint-2
	Data management	USN-6	As a user, I can enter data into the website securely	I can enter data only within the constraints	High	Sprint 1
		USN-7	As a user, I should give parameters of water as inputs	I can enter data only within the constraints	High	Sprint 1

6. PROJECT PLANNING AND SCHEDULING

a. . SPRINT PLANNING AND ESTIMATION

Project-Tracker:

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint1	Data Collection	USN-1,2	Collecting/ downloading dataset for pre-processing.	12	High	Kalanithi M Giridharan N
Sprint1	Data Preprocessing	USN-1,2	formats the data and handles the missing data in the dataset.	8	Medium	Karthick Raja R Bhuvaneshwar S
Sprint2	Model Building	USN-1,2	Calculate the Water Quality Index (WQI)using specified formula for every parameter.	10	High	Kalanithi M Bhuvaneshwar S
Sprint2	Accessing datasets	USN-1,2	Splitting the data into training and testing dataset from the entire dataset.	10	High	Karthick Raja R Giridharan N
Sprint3	Training and Testing	USN-1,2	Training the model using Random Forest Regression algorithm and testing the performance of the model (accuracy rate)	20	High	Giridharan N Bhuvaneshwar S
Sprint4	Implementation of Web page and user login	USN-1,2	Implementing the web page for collecting the data from user	12	High	Kalanithi M Bhuvaneshwar S
Sprint4	Web application	USN-1,2	It will display the current informationof the water quality.	8	Medium	Kalanithi M Karthick Raja R

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date	Story Points Completed	Sprint Release Date
Sprint1	20	6 Days	24 Oct 2022	29 Oct 2022	20	29 Oct 2022
Sprint2	20	6 Days	31 Oct 2022	05 Nov 2022	20	05 Nov 2022
Sprint3	20	6 Days	07 Nov 2022	12 Nov 2022	20	12 Nov 2022
Sprint4	20	6 Days	14 Nov 2022	19 Nov 2022	20	19 Nov 2022

Velocity:

Imagine we have a 10 days sprint duration and the velocity of the team is 20 (points per sprint). Let's cal team's average velocity AV per iteration unit.

Average Velocity:

Sprint 1 Average Velocity:

Average Velocity = 20/6 = 3.3

Sprint 2 Average Velocity:

Average Velocity = 20/6 = 3.3

Sprint 3 Average Velocity:

Average Velocity = 20/6 = 3.3

Sprint 4 Average Velocity:

Average Velocity = 20/6 = 3.3

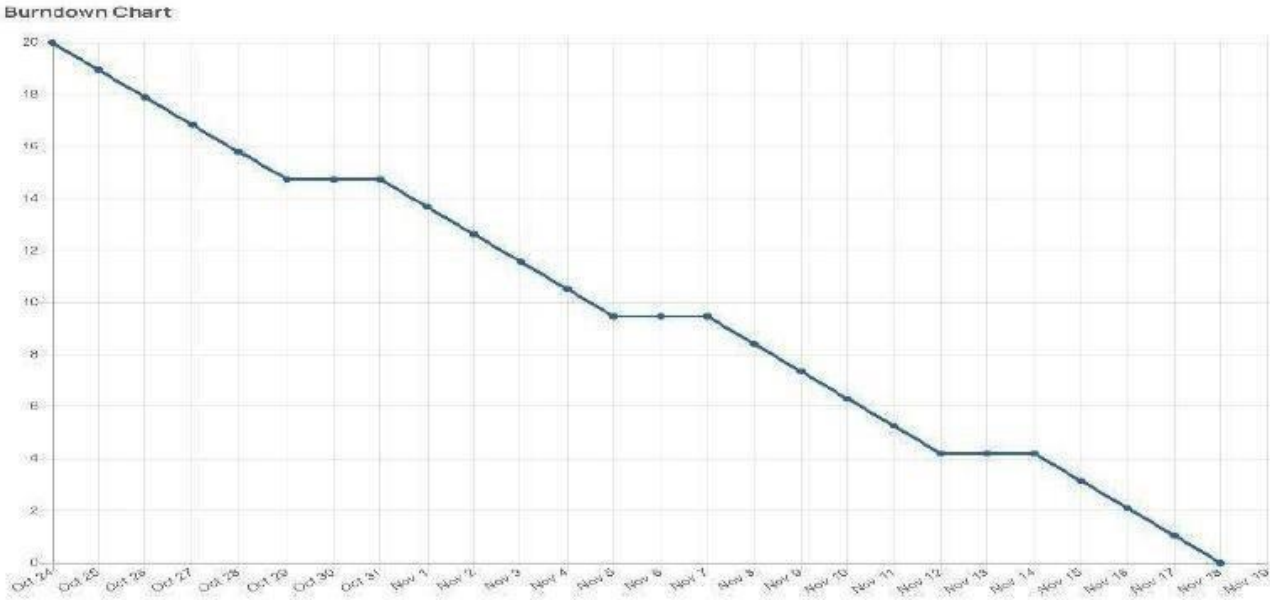
b. PROJECT DELIVERY SCHEDULE

TITLE	DESCRIPTION	DATE
Literature Survey & Information Gathering	Literature survey on the selected project & gathering information by referring the technical papers, research publications, journals etc.	01-10-2022
Prepare Empathy Map	Prepare Empathy Map Canvas to capture the user Pains and Gains, prepare list of problem Statements that are to be solved by this project.	02-10-2022
Ideation	List the ideas by organizing a brainstorming session and prioritize the top three ideas based on the feasibility and importance.	02-10-2022
Proposed Solution	Prepare the proposed solution document, which includes novelty, feasibility of idea, revenue model, social impact, scalability of solution, etc.	12-10-2022
Problem Solution Fit	Prepare problem - solution fit document.	30-10-2022
Solution Architecture	Prepare solution architecture document.	30-10-2022
Customer Journey	Prepare the customer journey maps to understand the user interactions and experiences with the application (entry to exit).	20-10-2022

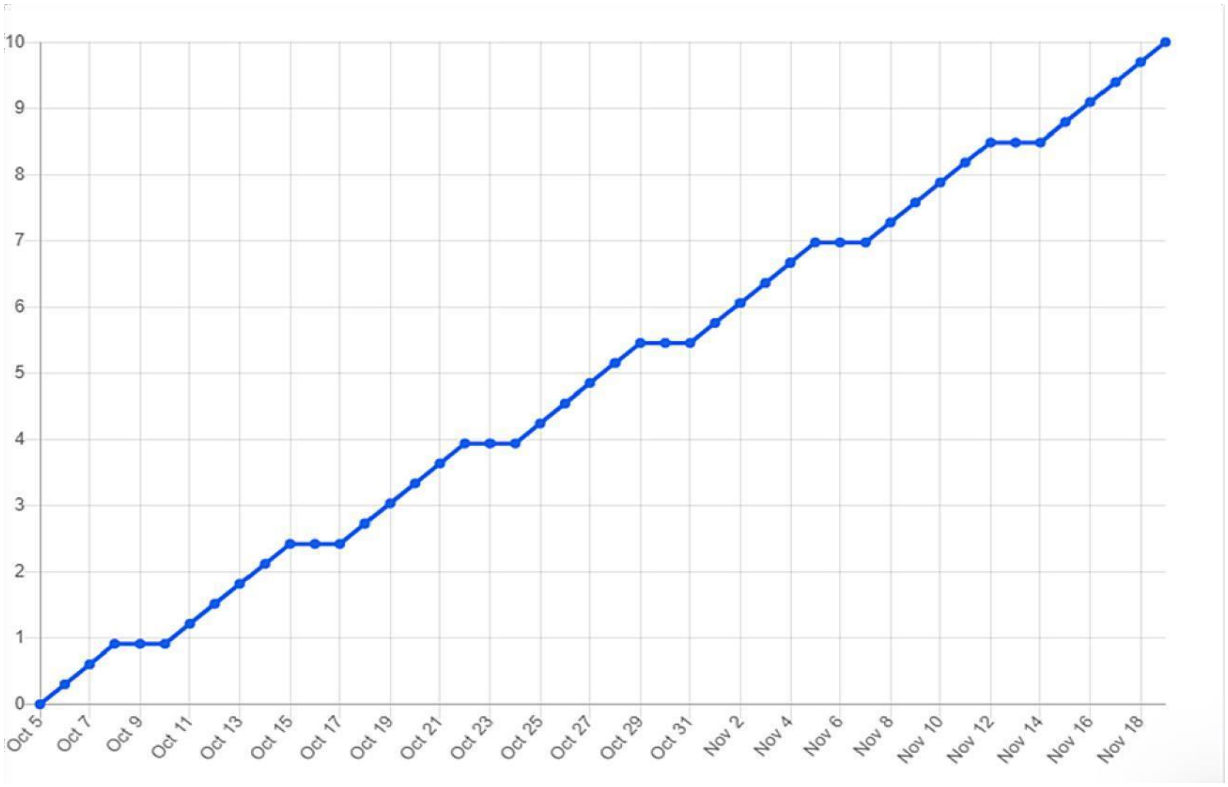
Functional Requirement	Prepare the functional requirement document.	11-10-2022
Data Flow Diagrams and User_Stories	Prepare the Data flow diagrams and User Stories for the problem	20-10-2022
Technology Stack Architecture	Prepare the Technology Stack Architechture	20-10-2022
Prepare Milestone &Activity List	Prepare the milestones and activity list of the project.	08-11-2022
Project Development Phase	Develop Project Development Phase which include sprint 1, sprint 2, sprint 3, sprint 4.	01-11-2022

c.REPORTS FROM JIRA

BURNDOWN CHART



BURN UP CHART



7. CODING AND SOLUTIONING

a. FEATURE 1 (RANDOMFOREST ALGORITHM MODEL)

Random Forest Classifier is used to train and test the model for detecting the Chronic Kidney Disease (CKD) with the help of collected and pre- processed dataset collections. NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high- level mathematical functions to operate on these arrays. Moreover, NumPy forms the foundation of the Machine Learning stack. Pandas is an open source Python package that

is most widely used for data science/data analysis and machine learning tasks. Sea born is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. For a brief introduction to the ideas behind the library, you can read the introductory notes or the paper. Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible. Create publication quality plots. Make interactive figures that can zoom, pan, update. EDA is applied to investigate the data and summarize the key in sights. It will give you the basic understanding of your data, it is distribution, null values and much more. You can either explore data using graphs or through some python functions. There will be two types of analysis. Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability. Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and Skewness. Label Encoding refers to converting the labels into a numeric form to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning. “Pickling” is the process whereby a Python object hierarchy is converted into a byte stream, and “unpickling” is the inverse operation, whereby a byte stream is converted back into an object hierarchy. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework.

b. FEATURE 2(FLASK CONNECTIVITY)

The framework is the basis upon which software programs are built.

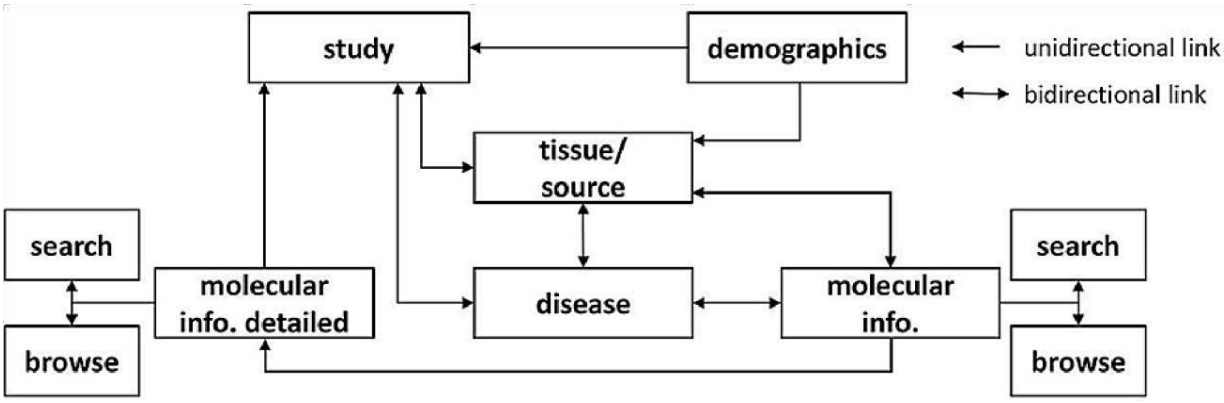
It serves as a foundation for software developers, allowing them to create a variety of applications for certain platforms. It is a set of functions and predefined classes used to connect with the system software and handle input sand outputs. It simplifies the life of a developer while giving them the ability to use certain extensions and makes the online applications scalable and maintainable. Flask is a web application framework written in Python. A Web Application Framework or a simply a Web Framework represents a collection of libraries and modules that enable web application developers to write applications without worrying about low-level details such as protocol, thread management, among other examples. Flask is a web application framework written in Python. It was developed by Armin Ronacher, who led a team of international Python enthusiasts called Pocco. Flask is based on the Werkzeg WSGI toolkit and the Jinja2 template engine. Both are Pocco projects. The Web Server Gateway Interface (Web Server Gateway Interface, WSGI) has been used as a standard for Python web application development. WSGI is the specification of a common interface between web server sand web applications.

Flask is often referred to as a micro-framework. It is designed to keep the core of the application simple and scalable. Instead of an abstraction layer for database support, Flask supports extensions to add such capabilities to the application. Unlike the Django framework, Flask is very Pythonic. It's easy to get started with Flask, because it doesn't have a huge learning curve. HTML stands for Hyper Text Markup Language. HTML is the standard markup language for creating Web pages. HTML describes the structure of a Web page. HTML consists of a series of elements. HTML elements tell the browser how to display the content. Flask is used for developing web applications using python, implemented on Werk zeug and Jinja2. Advantages of using Flask framework are: There is a built-in development server and a fast debugger provided. The model deployed using Flask is used to predict the Chronic Kidney Disease. Hypertext markup language (HTML) is the basic language used to create documents for the Web and, along with HTTP (hypertext transfer protocol) and URLs (universal resource locators), is

one of the three main protocols of the Web. Hypertext is text that contains hyperlinks. A hyperlink is an automated cross reference to another location on the same document or to another document which, when selected by a user, causes the computer to display the linked location or document within a concise period. A markup language is a set of tags that can be embedded in digital text to provide additional information about it, including its content, structure and appearance. This information facilitates automated operations on the text, including formatting it for display, searching it and even modifying it. Some type of markup language is employed by every word processing program and by nearly every other program that displays text, although such languages and their tags are typically hidden from the user.HTML consists of a set of predefined tags that can be embedded in text by web site designers in order to indicate the details of how web pages are rendered (i.e., converted into a final, easily usable, form) by web browsers. These details include paragraphing, margins, fonts (including style and size), columns, colors (background and text), links, the location of images, text flow around images, tables, and user input form elements

(such as spaces for adding text and submit buttons).

c. DATABASE SCHEMA



In the recent decades, the evolution of omics technologies has led to advances biological fields, creating a demand for effective storage, management and exchange of rapidly generated data and research discoveries. To address this need, the development of databases of experimental outputs has become a common part of scientific practice in order to serve as knowledge sources and data-sharing platforms, providing information about genes, transcripts, proteins or metabolites. In this review, we presentomics databases available currently, with a special focus on their application in kidney research and possibly in clinical practice. Databases are divided into two categories: general databases with a broad information scope and kidney- specific databases distinctively concentrated on kidney pathologies. In research, databases can be used as a rich source of information about pathophysiological mechanisms and molecular targets. In the future, databases will support clinicians with their decisions, providing better and faster diagnoses and setting the direction towards more preventive, personalized medicine. We also provide a test case demonstrating the potential of biological databases in comparing multi- omics datasets and generating new hypotheses to answer a critical and common diagnostic problem in nephrology practice. In the future, employment of databases combined with data integration and data mining should provide powerful insights into unlocking the mysteries of kidney disease, leading to a potential impact on pharmacological intervention and therapeutic disease management.

8. TESTING

a. TEST CASES

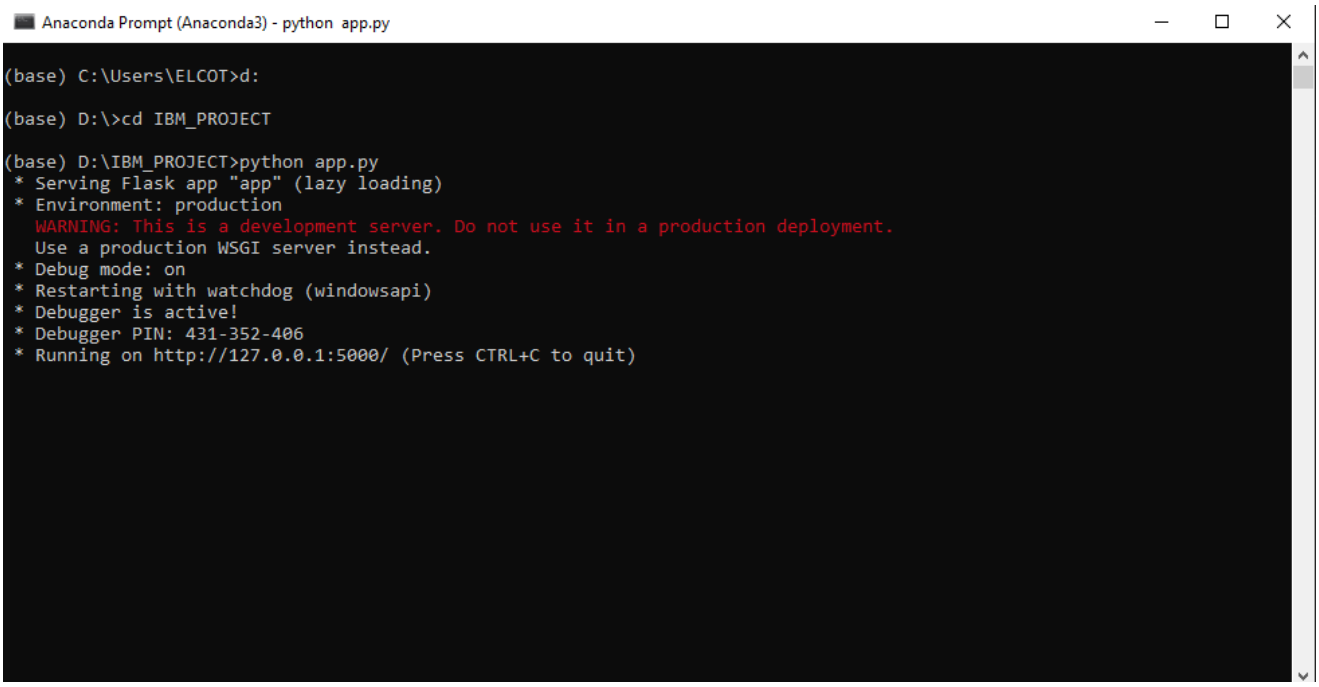
Test Scenario

Verify whether the deployed project predict sas per expected

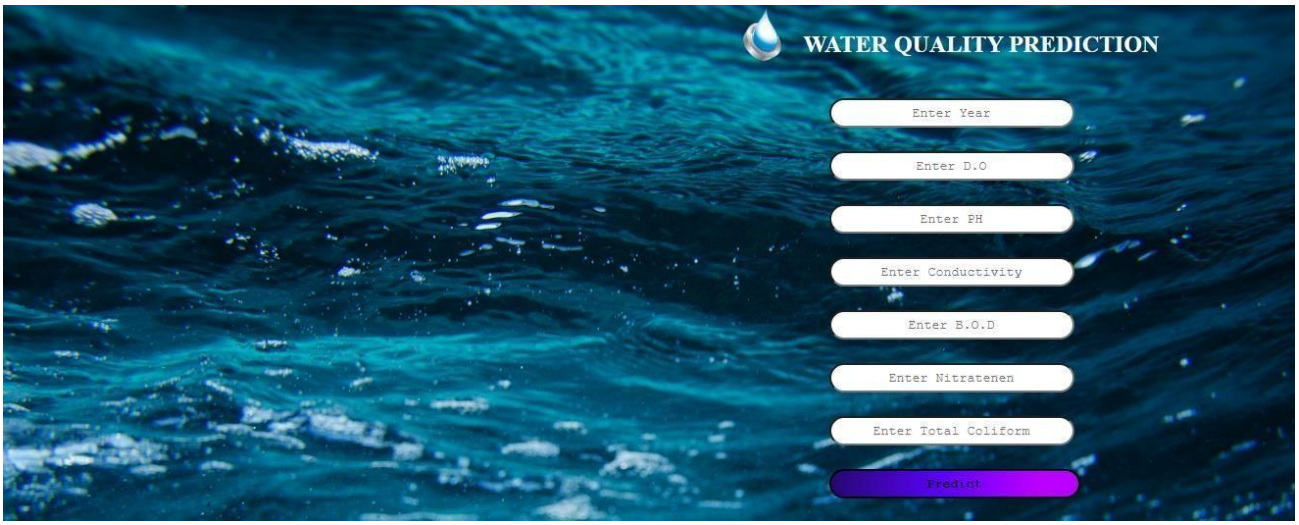
Step #	Step Details	Expected Results	Actual Results	Pass / Fail / Not executed / Suspended
1	Navigate to corresponding website link	Site should open	As Expected	Pass
2	Enter the details	Details should be entered	As Expected	Pass
3	Click Submit	Check the result	As Expected	Pass
4	Output results	Results are generated	As Expected	Pass

USER ACCEPTANCE TESTING

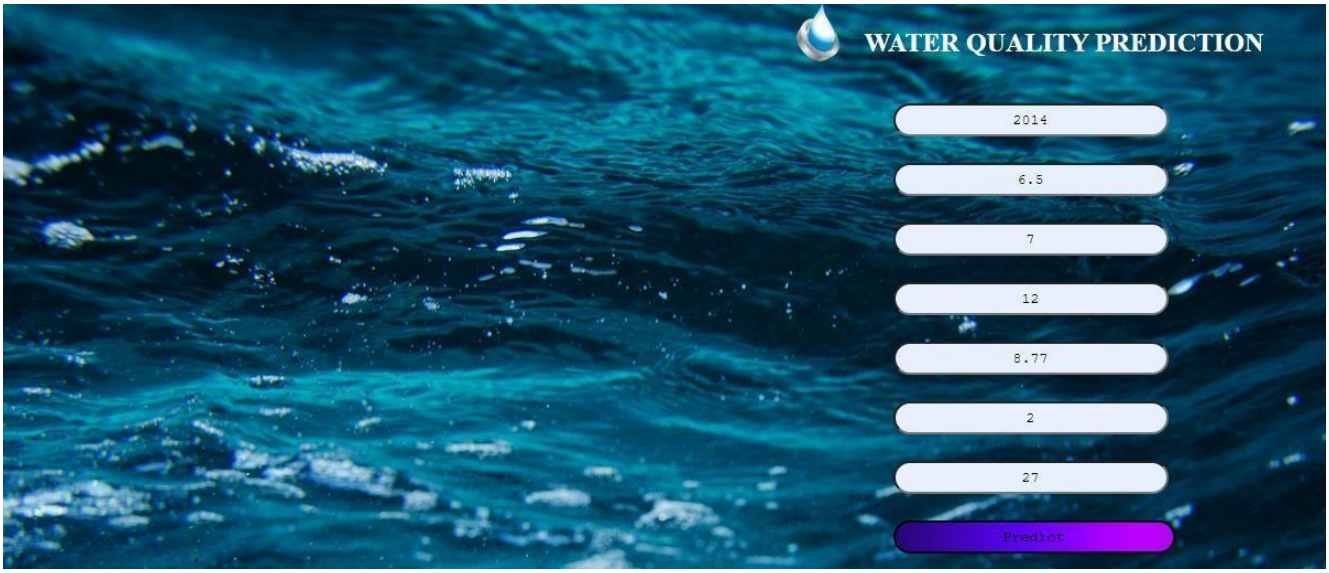
TEST CASE 1:


Anaconda Prompt (Anaconda3) - python app.py
(base) C:\Users\ELCOT>d:
(base) D:\>cd IBM_PROJECT
(base) D:\IBM_PROJECT>python app.py
* Serving Flask app "app" (lazy loading)
* Environment: production
WARNING: This is a development server. Do not use it in a production deployment.
Use a production WSGI server instead.
* Debug mode: on
* Restarting with watchdog (windowsapi)
* Debugger is active!
* Debugger PIN: 431-352-406
* Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)

TEST CASE 2:


WATER QUALITY PREDICTION
Enter Year
Enter D.O
Enter PH
Enter Conductivity
Enter B.O.D
Enter Nitratenen
Enter Total Coliform
Predict

TEST CASE 3:



9.RESULTS

a. PERFORMANCE METRICES

TITLE	DESCRIPTION	DATE
Literature Survey & Information Gathering	Literature survey on the selected project & gathering information by referring the technical papers, research publications, journals etc.	01-10-2022
Prepare Empathy Map	Prepare Empathy Map Canvas to capture the user Pains and Gains, prepare list of problem Statements that are to be solved by this project.	02-10-2022
Ideation	List the ideas by organizing a brainstorming session and prioritize the top three ideas based on the feasibility and importance.	02-10-2022
Proposed Solution	Prepare the proposed solution document, which includes novelty, feasibility of idea, revenue model, social impact, scalability of solution, etc.	12-10-2022
Problem Solution Fit	Prepare problem - solution fit document.	30-10-2022
Solution Architecture	Prepare solution architecture document.	30-10-2022

Customer Journey	Prepare the customer journey maps to understand the user interactions and experiences with the application (entry to exit).	20-10-2022
Functional Requirement	Prepare the functional requirement document.	11-10-2022
Data Flow Diagrams and User_Stories	Prepare the Data flow diagrams and User Stories for the problem	20-10-2022
Technology Stack Architecture	Prepare the Technology Stack Architechture	20-10-2022
Prepare Milestone &Activity List	Prepare the milestones and activity list of the project.	08-11-2022
Project Development Phase	Develop Project Development Phase which include sprint 1, sprint 2, sprint 3, sprint 4.	01-11-2022

10. ADVANTAGES AND DISADVANTAGES

ADVANTAGES:

Whether it be for groundwater, surface water or open water, there are a number of reasons why it is important for you to undertake regular water quality testing. If you're wanting to create a solid foundation on which to build a broader water management plan, then investing in water quality testing should be your first point of action. This testing will also allow you to adhere to strict permit regulations and be in compliance with Australian laws. Identifying the health of your water will help you to discover where it

may need some help. Ultimately, finding a source of pollution, or remaining proactive with your monitoring will enable you to save money in the long term. The more information that you can obtain will assist you with your decision on what product you may need to improve the condition of your water. Simply guessing and buying products based on a hunch or a general trend is ill-advised, as each body of water has unique properties that can only be discovered through testing. Measuring the amount of dissolved oxygen in your water is another important advantage of water quality testing, as typically the less oxygen, the higher the water temperature, resulting in a more harmful environment for aquatic life. These levels do fluctuate slightly across theseasons, but regular monitoring of your water quality will allow you to discover trends over time, and whether there are other factors that may be contributing to the results you discover.

DISADVANTAGES

Training necessary Some what difficult to manage over time and with large data sets Requires manual operation to submit data, some configuration required Costly, usually only feasible under Exchange Network grants Technical expertise and network server required Requires manual operation to submit data Cannot respond to dataqueries from other nodes, and therefore cannot interact with the Exchange Network Technical expertise and network server required.

11. CONCLUSION

Water is one of the most essential resources for survival and its quality is determined through WQI. Conventionally, to test water quality, one has to go through expensive and cumbersome lab analysis. This research explored an alternative method of machine learning to predict water quality using minimal and easily available water quality

parameters. The data used to conduct the study were acquired from PCRWR and contained 663 samples from 12 different sources of Rawal Lake, Pakistan. A set of representative supervised machine learning algorithms were employed to estimate WQI. This showed that polynomial regression with a degree of 2, and gradient boosting, with a learning rate of 0.1, outperformed other regression algorithms by predicting WQI most efficiently, while MLP with a configuration of (3, 7) outperformed other classification algorithms by classifying WQC most efficiently. In this paper, the performance of artificial intelligence techniques were evaluated to predict the water quality components of Tireh River (Iran). To this end most dataset related well known components, such as pH, SO_4, Na, Ca, Cl, Mg, HCO_3 etc., were collected. Results indicated that the applied models have suitable performance for predicting water quality.

12. FUTURE SCOPE

In future works, we propose integrating the findings of this research in a large-scale IoT-based online monitoring system using only the sensors of the required parameters. The tested algorithms would predict the water quality immediately based on the real-time data fed from the IoT system. The proposed IoT system would employ the parameter sensors of pH, turbidity, temperature and TDS for parameter readings and communicate those readings using an Arduino microcontroller and ZigBee transceiver. It would identify poor quality water before it is released for consumption and alert concerned authorities. It will hopefully result in curtailment of people consuming poor quality water and consequently deescalate harrowing diseases like typhoid and diarrhea. In this regard, the application of a prescriptive analysis from the expected values would lead to future facilities to support decision and policy makers. More data sources are required to verify the reliability and robustness of the proposed models. So far, the water quality dataset from the LVW collected by Southern Nevada Water Authority and Las Vegas Wash Coordination Committee, and dataset collected from Boulder Basin have been used as the experimental

dataset. In the future, more efforts will be made to find more datasets to build a more reliable water quality prediction model.

13. APPENDIX

Machine learning has been widely used as a powerful tool to solve problems in the water environment because it can be applied to predict water quality, optimize water resource allocation, manage water resource shortages, etc. Despite this, several challenges remain in fully applying machine learning approaches in this field to evaluate water quality:

1. Machine learning is usually dependent on large amounts of high quality data. Obtaining sufficient data with high accuracy in water treatment and management systems is often difficult owing to the cost or technology limitations.

2. As the conditions in real water treatment and management systems can be extremely complex, the current algorithms may only be applied to specific systems, which hinders the wide application of machine learning approaches.

3. The implementation of machine learning algorithms in practical applications requires researchers to have certain professional background knowledge.

To overcome the above-mentioned challenges, the following aspects should be considered in future research and engineering practices:

1. More advanced sensors, including soft sensors, should be developed and applied in water quality monitoring to collect sufficiently accurate data to facilitate the application of machine learning approaches.

2. The feasibility and reliability of the algorithms should be improved, and more universal algorithms and models should be developed according to the water treatment and management requirements.

3. Interdisciplinary talent with knowledge in different fields should be trained to develop more advanced machine learning techniques and apply them in engineering practices.

REQUIREMENT.TXT

Flask = 2.2.2

Joblib = 1.2.0

Numpy = 1.23.4

Pandas =1.5.1

Scikit-learn =1.1.3

Xgboost = 1.7.1

Gunicorn= 20.1.0

Matplotlib = 3.6.2

Seaborn = 0.12.1

Source Code

App.py

```
import numpy as numpy
from flask import Flask,render_template,request
import pickle


app = Flask(__name__)
model = pickle.load(open('wqi.pkl','rb'))
@app.route('/',methods=['GET'])
def home() :
 return render_template("web.html")

@app.route('/login',methods = ['POST'])
def login() :
    year = request.form['year']
    do = request.form["do"]
    ph = request.form["ph"]
    co = request.form["co"]
    bod = request.form["bod"]
    na = request.form["na"]
    tc = request.form["tc"]
    total = [[int(year),float(do),float(ph),float(co),float(bod),float(na),float(tc)]]
    y_pred = model.predict(total)
    y_pred = y_pred[[0]]
    if(y_pred >= 95 and y_pred<=100) :
        return render_template("web.html",showcase = 'Excellent,The predicted value is '+ str(y_pred))
    elif(y_pred >= 89 and y_pred <= 94) :
        return render_template("web.html",showcase = 'Very good,The predicted value is '+ str(y_pred))
    elif(y_pred >= 80 and y_pred <= 88) :
        return render_template("web.html",showcase = 'Good,The predicted value is '+ str(y_pred))
    elif(y_pred >= 65 and y_pred <= 79) :
        return render_template("web.html",showcase = 'Fair,The predicted value is '+ str(y_pred))
    elif(y_pred >= 45 and y_pred <= 64) :
        return render_template("web.html",showcase = 'Mariginal,The predicted value is '+ str(y_pred))
    else :
        return render_template("web.html",showcase = 'Poor,The predicted value is '+ str(y_pred))


    if__name__== '_main_' :
    app.run(debug = True,port=5000)
```

Web.html

```
<!DOCTYPE html>
<html lang="en">

<head>
   <meta charset="UTF-8">
   <meta http-equiv="X-UA-Compatible" content="IE=edge">
   <meta name="viewport" content="width=device-width, initial-scale=1.0">
   <title>Water quality prediction</title>
   <link rel="stylesheet" href="../static/css/style.css">

</head>

<body>
   <header>
                  <nav>
         <div class="row">
            <div class="row1">
               <img src="../static/css/water_droplet.png" alt="logo">
                                 </div>
            <div class="row2">
               <h1>WATER QUALITY PREDICTION</h1>
            </div>
         </div>
      </nav>
   </header>
   <main>
      <div class="column">
         <form action="/login" method="post">
            <label for=""></label>
            <input type="text" name="year" id="" placeholder="Enter Year">
            <label for=""></label>
            <input type="text" name="do" id="" placeholder="Enter D.O">
            <label for=""></label>
            <input type="text" name="ph" id="" placeholder="Enter PH">
            <label for=""></label>
            <input type="text" name="co" id="" placeholder="Enter Conductivity">
            <label for=""></label>
            <input type="text" name="bod" id="" placeholder="Enter B.O.D">
            <label for=""></label>
            <input type="text" name="na" id="" placeholder="Enter Nitratenen">
            <label for=""></label>
            <input type="text" name="tc" id="" placeholder="Enter Total Coliform">
```

```
            <label for=""></label>
            <div class="last">
                <input type="submit" value="Predict">
            </div>
            <div class="bor">
                {{showcase}}
            </div>

        </form>
    </div>
  </main>
  </div>
</body>

</html>
```

LINKS:

GITHUB :
https://github.com/IBM-EPBL/IBM-Project-47946-1660803505

IBM-Project-47946-1660803505

PROJECT DEMO:

https://youtu.be/YUFF8WkNdrs